Computers & Graphics 35 (2011) 54–66

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Extended papers from NPAR 2010

Stylized ambient displays of digital media collections

Tinghuai Wang [a,*,1], John Collomosse [a], Rui Hu [a], David Slatter [b], Darryl Greig [b], Phil Cheatle [b]

[a] Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[b] Multimedia Interaction and Understanding Lab, Hewlett-Packard Labs, Bristol, UK

ARTICLE INFO

Article history:
Received 30 August 2010
Received in revised form
1 November 2010
Accepted 4 November 2010
Available online 18 November 2010

Keywords:
Artistic rendering
Segmentation
Graph cut
Temporal coherence
Composition
Ambient displays

ABSTRACT

The falling cost of digital cameras and camcorders has encouraged the creation of massive collections of personal digital media. However, once captured, this media is infrequently accessed and often lies dormant on users' PCs. We present a system to breathe life into home digital media collections, drawing upon artistic stylization to create a "Digital Ambient Display" that automatically selects, stylizes and transitions between digital contents in a semantically meaningful sequence. We present a novel algorithm based on multi-label graph cut for segmenting video into temporally coherent region maps. These maps are used to both stylize video into cartoons and paintings, and measure visual similarity between frames for smooth sequence transitions. The system automatically structures the media collection into a hierarchical representation based on visual content and semantics. Graph optimization is applied to adaptively sequence content for display in a coarse-to-fine manner, driven by user attention level (detected in real-time by a webcam). Our system is deployed on embedded hardware in the form of a compact digital photo frame. We demonstrate coherent segmentation and stylization over a variety of home videos and photos. We evaluate our media sequencing algorithm via a small-scale user study, indicating that our adaptive display conveys a more compelling media consumption experience than simple linear "slide-shows".

1. Introduction

Traditional approaches to linearly browsing personal media collections (e.g. photo albums, slideshows) are becoming impractical due to the explosive growth of media repositories. Furthermore, although digital media is intrinsically more accessible than physical archives, the focus on the PC as the main portal to these collections poses a convenience barrier to realizing their value. The proliferation of video and image data in digital form creates demand for an effective means to browse large volumes of digital media in a structured, accessible and intuitive manner.

This paper proposes a novel approach to the consumption of home digital media collections, centred upon *ambient experiences*. Ambient experiences are distinguished from compelling or intense experiences in that they are able to co-exist harmoniously with other activities such as conversations, shared meals and so forth. An ambient experience does not demand the full attention of the user but is able to play out in a pleasing, unobtrusive way such that fresh and interesting content is available in the attention spaces of everyday life.

People have been creating ambient displays of media for decades; by placing photographs on the mantelpiece, or leaving the television or radio on in the background. These established behaviors reflect commonly appreciated value in the passive consumption of media. Yet, beyond updates to accommodate new capture technology (e.g. digital photo frames) methods of passive content delivery and consumption have remained largely unchanged. Our work aims to create a platform for ambient delivery of home digital media content. Such content often encodes householders' memories and life experiences. We seek to emulate, for digital media, the serendipitous process of rediscovery often experienced whilst browsing physical media archives (e.g. a box of photos in the attic) that can trigger enjoyable reminiscence over past memories and events.

Although considerable research has been devoted to direct interactive approaches for browsing digital media collections, there is little previous work addressing the problem of displaying digital content in an *ambient* manner (Section 1.2). Typical approaches to browsing small or medium scale photo sets project thumbnails onto either a planar or a spherical surface, so that images that are visually similar are located in close proximity in the visualization [1–4]. Large photo sets are often handled by clustering content into subsets, sometimes arranged hierarchically for visualization and manual navigation [5–7]. With the expected proliferation of large format video displays around the home, recent work explores the specific domain of household digital media interaction [8,9]. Yet, the *ambient* dissemination of home

* Corresponding author.
E-mail addresses: tinghuai.wang@surrey.ac.uk (T. Wang), j.collomosse@surrey.ac.uk (J. Collomosse), r.hu@surrey.ac.uk (R. Hu), david.slatter@hp.com (D. Slatter), darryl.greig@hp.com (D. Greig), phil.cheatle@hp.com (P. Cheatle).

[1] Funded by the Hewlett-Packard Labs IRP Programme.

0097-8493/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cag.2010.11.004

visual media, and the associated issues of interaction in an ambient context, remain sparsely researched.

An earlier version of this extended paper [10] presented a video-only digital ambient display with simple media sequencing. The novel contribution of this paper over [10] is a more sophisticated approach to sequencing, involving unsupervised hierarchical clustering of media and content navigation influenced by user attention. In addition, an implementation on embedded hardware has enabled an in situ user evaluation of the digital ambient display, also presented in this paper.

1.1. Digital ambient display (DAD) concept

The *digital ambient display* (DAD) is an always-on display for living spaces that enables users to effortlessly visualize and rediscover their personal digital media collections. DADs address the paradoxical requirement of an autonomous technology to passively disseminate media collections, that also enables minimal interaction to actively navigate routes through content that may trigger interest and user reminiscence. By transitioning between selected media items, the DAD passively presents a global summary visualizing the essential structure of the collection. This results in an evolving temporal composition of media, the sequencing of which considers both media semantics and visual appearance, as well as adaptively responding to user attention (sensed via gaze detection). Rather than simply stitching digital content together, we harness artistic stylization to depict image and video in a more abstract sense. In contrast to photorealism, which often proves distracting in the ambient setting (e.g. a television in the corner of a café), artistic stylization would provide an aesthetically pleasing and unobtrusive means of disseminating content in the ambient setting; creating a flowing, temporal composition that conveys the essence of users' experiences through an artistic representation of their digital media collection (Fig. 1).

Creating a DAD requires that the media be automatically parsed into an structured representation that enables semantically meaningful routes to be navigated through the collection. This process is dependent on meta-data tags user-assigned to each media item. Furthermore, the visual content within individual media items must be also be parsed into a mid-level *visual scene representation* that enables both:

1. Artistic rendering of media into aesthetically pleasing forms.
2. Generation of appropriate transition effects and sequencing decisions, to create an appealing temporal composition.

While artistic rendering of images is much explored, temporally coherent stylization of video is still a challenging task which requires a stable and consistent description of the scene structures. Following DeCarlo and Santella [11] as well as Collomosse et al. [12], we identify a color region segmentation as being an appropriate "mid-level" scene abstraction, and in Section 4 contribute a novel algorithm for segmenting video frames into a deforming set of temporally coherent regions. We demonstrate how these regions may be stylized via either shading or stroke-based rendering, to produce coherent cartoon and painterly video styles (Section 4.5). We describe our hierarchical approach to structuring the media collection in Section 3, and describe how content is sequenced at run-time using that representation in Section 5. Qualitative evaluations of segmentation coherence, and a quantitative user evaluation of the DAD, are presented in Section 6.

1.2. Related work

Video temporal composition was first proposed by Schodl et al. [13], within the scope of a single video and based upon visual similarity only. Much as motion graphs [14] construct a directed graph that encapsulates connections among motion capture fragments, so the Video Textures of Schodl et al. create a graph of video fragments that may be walked in perpetuity to create a temporal composition. Our proposed approach to composition borrows from the graph representations [14,13], but using a *hierarchical* representation, comprising multiple videos, and measuring similarity both visually and semantically.

Hierarchical structuring of media is common to many contemporary approaches for interactively navigating collections. For example, Krishnamachari et al. form tree structures from an image collection, imposing a coarse to fine representation of image content within clusters and enabling the users to navigate up and down the tree levels via representative images from each cluster. This approach was later adopted by Chen et al. [5] and Goldberger et al. [6]. Chen et al. propose a fast search algorithm and a fast-sparse clustering method for building hierarchical tree structures from large image collections. Goldberger et al. combine discrete and continuous image models with information-theoretic based criteria for unsupervised hierarchical clustering. Images are clustered such that the mutual information between the clusters and the image content is maximally preserved. Our approach to structuring collections combines a hierarchical clustering with graph optimization approach [14] to navigate and visualize large media collections. The resulting system differs from existing hierarchical clustering approaches [5–7] in several ways. Rather than exploiting low level visual features in the clustering, we incorporate both high-level semantic similarity when constructing top levels of the tree, and global image feature descriptors via a Bag of visual words (BoW) framework [15] for constructing lower levels of the tree. Consequently, this tree structure not only enables a global semantic summary of the collection, but also encodes visual similarities at various levels. Furthermore, in our system each node in the hierarchy encodes a directed graph that encapsulates connections among the digital items assigned to thats node, rather than an unstructured subset of media as typified by previous work.

Video stylization was first addressed by Litwinowicz [16], who produces painterly video by pushing brush strokes from frame to frame in the direction of optical flow motion vectors. This approach was later extended by Hayes and Essa [17] who similarly move strokes but within independent motion layers. Complementary work by Hertzmann [18] use differences between consecutive frames of video, painting over areas of the new frame that differ significantly from the previous frame. While these methods can

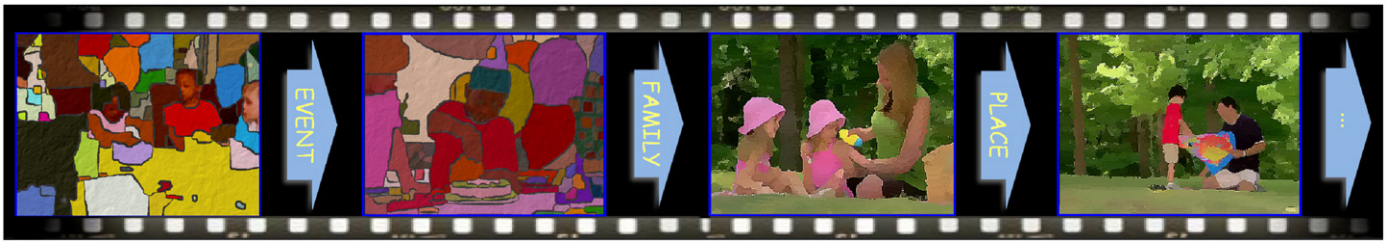

Fig. 1. The digital ambient display (DAD) selects, stylizes and transitions between home digital media items (photos and videos) according to semantic and visual similarity. The paths through the media collection are passively influenced by user interest measurement (gaze detection).

produce impressive painterly video, the errors in the estimated per-pixel motion field can quickly accumulate and propagate to subsequent frames, resulting in increasing temporal incoherence. This can lead to a distraction scintillation or "flicker" when strokes of the stylized output no longer match object motion [19].

More recently, image segmentation techniques have been applied to yield *mid-level* models of scene structure [20,21] that can be rendered in artistic styles. By extending the mean-shift based stylization approach on images [11], Collomosse et al. [12] create spatio-temporal volumes from video by associating 2D segmentations over time and fitting *stroke surfaces* to voxel objects. Although this geometric smoothing improves stability, temporal coherence is not ensured because the region map for each frame is formed independently without knowledge of the adjacent frames. Furthermore, association is confounded by the poor repeatability of 2D segmentation algorithms between similar frames, causing variations in the shape and photometric properties of regions that require manual correction. Wang et al. [21] also transform video into spatio-temporal volumes by clustering space-time pixels using a mean-shift operator. However, this approach becomes computationally infeasible for pixel counts in even moderate size videos, and often under-segments small or fast moving objects that form disconnected volumes. This also requires manual correction and frequent grouping of space-time volumes.

Winnemoeller et al. [22] present a method to abstract video using a bilateral filter, attenuating detail in low-contrast regions using an approximation to anisotropic diffusion while artificially increasing contrast in higher contrast regions with difference-of-Gaussian edges. A variant of the bilateral filter is presented by Kyprianidis et al. [23] which is aligned to the local structure of the image to avoid block artifacts and create smooth boundaries between color regions. Another variant of the bilateral filter is presented by Kang et al. [24] that is guided by a feature flow field, which improves feature preservation, noise reduction, and stylization. A further technique based on a generalization of the Kuwahara filter has been presented by Kyprianidis et al. [25] using the Kuwahara filter. Bhat et al. [26] provide a solution for applying gradient-domain filters to videos and video streams in a coherent manner. Such approaches do not seek to parse a description of scene structure, making them useful for scenes that are difficult to segment, but limited to a characteristic soft-shaded artistic style.

We adopt a scene segmentation approach; convenient both for diverse artistic rendering and for creating the structural correspondences between frames for transition animations. We propose a new video segmentation algorithm, in which the segmentation of each frame is guided by motion flow propagated priors estimated from the region labels of past frames. In doing so we combine the automation of early optical flow stylization algorithms with the robustness and coherence of region segmentation approaches; propagating labels with flow, and resolving ambiguities using a graph-cut optimization to create coherent region maps. Some recent interactive "video cut-out" systems are similar in spirit [27,28]; tracking key-points on region boundaries over time for matte segmentation. However we differ in several ways. First, we propagate label priors and data forward with motion flow within regions, rather than tracking 2D windows on region boundaries that contain clutter from adjacent regions. Second, we are more general, producing a multi-label (region) map rather than a binary matte. Third, both interactive systems [27,28] require regular manual correction, typically every ~ 5 frames. Our algorithm requires no user interaction, beyond (optional) modification of the initial frame for aesthetics.

2. System overview

The Digital Ambient Display (DAD) visualizes home media collections comprising photos and videos. Videos are ingested as short, visually interesting clips that form the atomic unit of composition. Obtaining such clips differs from classical shot detection as raw home footage tends to consist of a few lengthy shots. An existing algorithm [29] performs this pre-processing. The ingested media collection is clustered into a hierarchical representation according to semantic content (derived from keyword meta-data tags attached to the photo or video), and visual similarity (computed from the photo, or a representative video frame—typically mid-sequence). The automated clustering and related pre-processing are described in Section 3.

At run-time the DAD creates a temporal composition of a subset of media items from the structured collection, creating a media sequence that flows smoothly with respect to both scene appearance and semantics. For example, the DAD might select a clip or image of the family in the garden, and follow this with a family clip or image in semantically similar alternate environment such as a park. To promote user interest in the display, media choice is also governed by the level of user attention; passively measured using gaze detection. Persistent attention will guide the temporal composition toward semantically similar content to that which attracted the user's gaze. This real-time sequencing process is described in Section 5.

Presentation of video in the DAD is underpinned by a novel algorithm for segmenting video frames into temporally coherent colored regions (Sections 4.1–4.3). These region maps form a stable representation of visual structure in the scene that is used both to drive artistic rendering algorithms for stylization (Section 4.5), and to perform matching of scene elements between frames in order to generate animated clip transitions (Section 5.2). Photographs are similarly segmented into colored regions, and for convenience are treated as single-frame videos within our framework. However several segmentation-based photo stylization algorithms exist and might trivially be applied in place of our approach. Fig. 2 provides an overview of the complete DAD system.

3. Structuring the media collection

We represent the media collection as a hierarchy of pointers to media items. Each node in the tree represents a subset of the media collection sharing a common semantic theme or visual appearance.

3.1. Hierarchical clustering

Our top-down approach recursively splits the collection, applying unsupervised clustering to each node using the Affinity Propagation (AP) algorithm [30]. In contrast to k-means clustering, which iteratively refines an initial randomly chosen set of exemplars, AP simultaneously considers all data points within a node as potential exemplars and iteratively exchanges messages between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. AP requires only a measure of similarity between items, rather than a feature vector, and does not require *prior* knowledge of the dataset such as the number of clusters present and representatives for the different clusters.

Clustering proceeds in two phases. The initial phase constructs higher levels of the tree using a measure of semantic similarity (Section 3.2) that exploits user-provided tags on media items. AP is applied recursively to each node until no further division occurs (i.e. no semantic differentiation can be made between media items at a particular node). The second phase of our process then constructs lower levels of the tree from the leaf nodes of the first phase, using AP to cluster items based on a measure of visual similarity (Section 3.3). Higher levels of the tree thus provide semantic summaries of the media reflecting the diversity of the visual content in the dataset. Clusters at lower levels contain predominantly visually similar images at various levels of detail (Fig. 3).

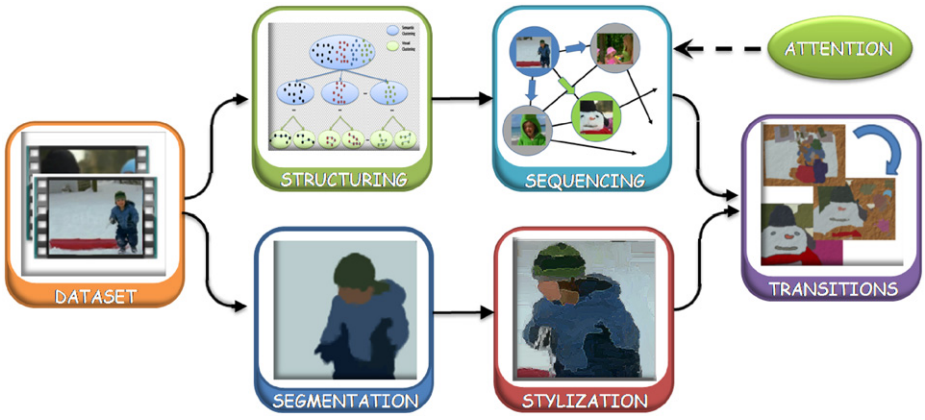

Fig. 2. System overview. User media (photos, videos) are hierarchically clustered according to content semantics and appearance. The sequencing of displayed content is driven both by this automated clustering, and via passive measurement of user attention. Videos and photos are segmented into region maps encoding visual structure. These region maps drive the artistic stylization and transition processes on the display.

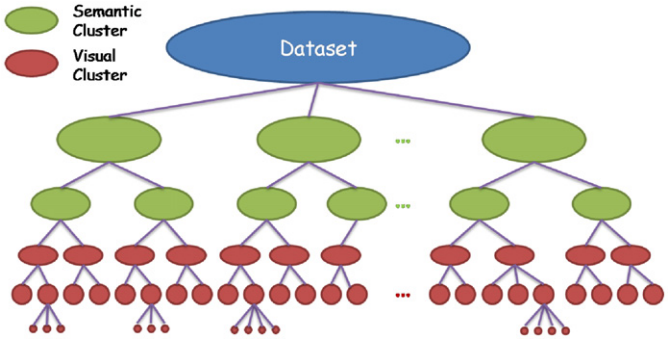

Fig. 3. The structuring process starts with the whole dataset corresponding to the root node of the tree and continues splitting until all the leaf nodes cannot be further split by recursively applying unsupervised clustering. Higher levels of the tree are clustered based on semantic (keyword) similarity, while lower levels are constructed based on visual similarity.

3.2. Semantic similarity

To compute the semantic similarity (S_s) of a pair of media items, we measure their tag co-occurrence. Given a vocabulary $\mathcal{V}=\{w_1,\ldots,w_K\}$ of K keywords present within all user-provided tags, the similarity of a pair of keywords is computed using asymmetric co-occurrence [31], indicating the probability of w_i appearing in a tag set given the presence of w_j:

$$p(w_i|w_j)=\frac{|w_i\cap w_j|}{|w_j|}, \tag{1}$$

where $|w_i\cap w_j|$ denotes the number of media item tagged with both keywords w_i and w_j, while $|w_j|$ is the number of items tagged with keyword w_j.

Following Sigurbjörnsson and van Zwol [31], we compute the asymmetric similarity between two sets of tags containing multiple keywords $T^1=\{w_1^1,w_2^1,\ldots,w_N^1\}$ and $T^2=\{w_1^2,w_2^2,\ldots,w_M^2\}$ with super-scripts corresponding to media items I_1 and I_2 respectively as

$$S_s(I_2|I_1)=\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}p(w_m^2|w_n^1)}{M\cdot N}. \tag{2}$$

Fig. 4 shows an example of tagged keywords and resulting semantic similarities between pairs of media items.

3.3. Visual similarity

We adopt a content-based image retrieval (CBIR) approach to compute the visual similarity (S_v) of two media items. Given a set of media items at a particular node, we adopt a *bag of visual words* (BoW) framework to create a codebook of visual words from discriminative features (descriptors) local to visual keypoints detected within each item.

Fig. 4. Tags of media items containing multiple keywords and semantic similarities based on tag co-occurrence.

Scale-invariant keypoints are obtained with the Harris–Laplace point detector [32] and then are described using the SIFT descriptor. Harris-SIFT descriptors from all images and key-frames extracted from video clips are clustered to form a BoW codebook via k-means clustering. A frequency histogram H^I is constructed for each I, indicating the visual words present within that media item. Visual similarity is then computed by measuring the histogram intersections of media pairs:

$$S(H^1,H^2)=\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_{ij}\cdot\min(H^1(i),H^2(j)),$$

$$\omega_{ij}=1-|\mathcal{H}^1(i)-\mathcal{H}^2(j)|, \tag{3}$$

where $H(i)$ indicates the ith bin of the histogram, $\mathcal{H}(i)$ the normalized visual word corresponding to the ith bin.

4. Video stylization

We next describe a coherent video segmentation algorithm which performs a multi-label graph cut on successive video frames, using both photometric properties of the current frame and prior information propagated forward from previous frames. This information comprises:

1. an incrementally built Gaussian Mixture Model (GMM) encoding the color distribution of each region over past frames;

2. a subset of pixel-to-region labels from the previous frame.

We check for region under-segmentation (e.g. the appearance of new objects, or objects emerging from occlusion) by comparing the historic and updated GMM color models for each region, and introducing new labels into that region if the color model appears to be temporally inconsistent. The region map of the first frame is boot-strapped using mean-shift segmentation [33], and may *optionally* be modified by the user for aesthetics e.g. to abstract away background detail by merging regions. Fig. 5 gives an overview of the segmentation algorithm.

We first describe our video segmentation algorithm (Sections 4.1–4.3) and then describe how the coherent region maps are applied to stylize video (Sections 4.4 and 4.5) and create the animations used to transition between successive clips in the DAD sequence.

4.1. Multi-label graph cut

We formulate segmentation as the problem of assigning region labels existing in frame I_{t-1} to each pixel $p \in \mathcal{P}$ in frame $I_t(p)$; i.e. seeking the best mapping $l : \mathcal{P} \to \mathcal{L}$ where $\mathcal{L} = (l(1), \ldots, l(p), \ldots, l(|\mathcal{P}|))$ is the set assignments of labels $l_i, i = \{1, \ldots, L\}$, and $\mathcal{P}$ is an 8-connected lattice of pixels.

A subset of $\mathcal{L}$ are carried forward from the region map at $t-1$, via a propagation process described shortly (Section 4.2). This *prior labeling* of pixels ($\mathcal{O} \subseteq \mathcal{P}$) forms a hard constraint on the assignments of remaining pixels in I_t, which are labeled to minimize a global energy function encouraging both temporal consistency of color distribution between frames, and spatial homogeneity of contrast within each frame. This is captured by the data and pairwise terms of the Gibbs energy function:

$$E(\mathcal{L}, \Theta, \mathcal{P}) = U(\mathcal{L}, \Theta, \mathcal{P}) + V(\mathcal{L}, \mathcal{P}). \tag{4}$$

The data term $U(.)$ exploits the fact that different color homogeneous regions tend to follow different color distributions. This encourages assignment of pixels to the labeled region following the most similar color model (we write the parameters of such models Θ). The data term is defined as

$$U(\mathcal{L}, \Theta, \mathcal{P}) = \sum_{p \in \mathcal{P}} -\log P_g(I_t(p) | l(p); \Theta),$$

$$P_g(I(p) | l(p) = l_i; \Theta) = \sum_{k=1}^{K_i} w_{ik} \mathcal{N}(I(p); \mu_{ik}, \Sigma_{ik}), \tag{5}$$

i.e. the data model of the ith label l_i is represented by a mixture of Gaussians (GMM), with parameters w_{ik}, μ_{ik} and Σ_{ik} representing the weight, the mean and the covariance of the kth component. The parameters of all GMMs ($\Theta = \{w_{ik}, \mu_{ik}, \Sigma_{ik}, i = 1, \ldots, L, k = 1, \ldots, K_i\}$) are learned from historical observations of each region's color distribution (Section 4.2).

The contrast term $V(\cdot)$ encourages coherence in region labeling and discontinuities to occur at high contrast locations, which is computed using RGB color distance as in GrabCut [34]:

$$V(\mathcal{L}, \mathcal{P}) = \gamma \sum_{(m,n) \in N} [l(m) \neq l(n)] e^{-\beta \|I(m) - I(n)\|^2}, \tag{6}$$

where N is the set of pairs of 8-connected neighboring pixels in $\mathcal{P}$. β is chosen to be contrast adaptive [35]:

$$\beta = \tfrac{1}{2} \langle \|I(m) - I(n)\|^2 \rangle^{-1}, \tag{7}$$

where $\langle \cdot \rangle$ denotes expectation over an image sample. Constant γ is a versatile setting for a variety of images [36], and is set empirically to obtain satisfactory segmentation.

Motivated by the data term defined by Boykov and Funka-Lea [35], we enforce hard constraints on the motion propagated prior labels assigned to label l_i, by setting the data term of $p \in \mathcal{O}$ to be

$$U_{p:\{p \in \mathcal{O}\}} = \begin{cases} 0 & \text{if } l(p) = l_i, \\ \infty & \text{if } l(p) \neq l_i. \end{cases} \tag{8}$$

Optimizing (4) to yield an appropriate assignment of labels to pixels is NP-hard, but an approximate solution can be computed by treating the optimization as a multi-label graph cut and solving this using the expansion move algorithm [37]. An $\alpha-$expansion iteration is a change of labeling such that p either retains its current value or takes the new label l_α. The expansion move proceeds by cycling the set of labels and performing an $\alpha-$expansion iteration for each label until (4) cannot be decreased [37]. Each $\alpha-$expansion iteration can be solved exactly by performing a single graph cut using the min-cut/max-flow [38]. Convergence to a strong local optimum is usually achieved in 3–4 cycles of iterations over our label set. We improve the computation and memory efficiency of each iteration by dynamically reusing the flow at each iteration of the min-cut/max-flow algorithm (after Alahari et al. [39]). This results in a $\times 2$ speed-up.

4.2. Region propagation

The segmentation of I_t described in Section 4.1 is dependent on the information propagated from the previous frame at $t-1$; specifically: (i) the color models for regions Θ; (ii) the set of pixels $\mathcal{O} \subseteq \mathcal{P}$ and their corresponding label assignments at $t-1$. We now explain the propagation process in detail.

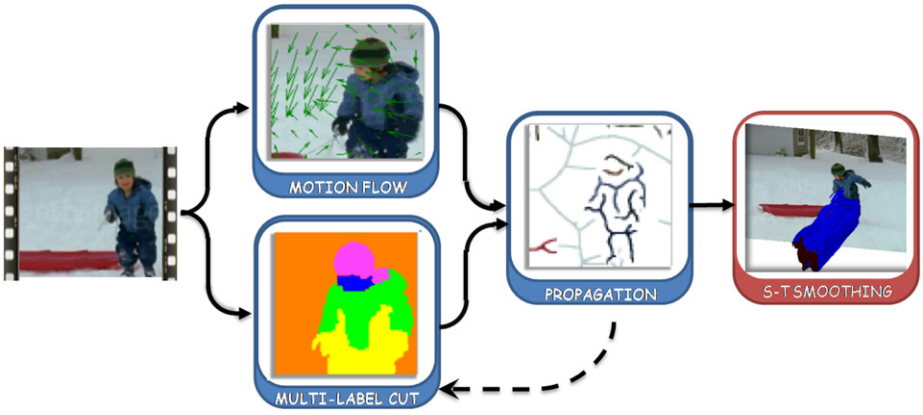

Fig. 5. Content stylization is underpinned by a graph-cut based segmentation of photos and video frames. In the case of video, region maps from previous frames act as a prior on the graph cut of successive frames. The skeleton of each region in the previous frame is propagated to the next frame using optical flow, along with an associated GMM color model, to form as constraints on the graph cut. New regions are detected via non-conformity with the propagated color models. The resulting region maps are smoothed via spatio-temporal low-pass filtering (after [12]).

Our approach is to estimate the motion of pixels in frame I_{t-1}, and translate those pixels and their respective label assignments from the previous frame to the current frame (I_t). Motion is estimated using a model of *rigid motion plus deformation*.

We first estimate a global affine transform between successive frames I_{t-1} and I_t, using a RANSAC search based on SIFT features [40] matched between the frames. Performing an affine warp on I_t and the corresponding region map compensates for large rigid (e.g. camera) motion, resulting in a new image I'_{t-1}. Local deformations are captured by estimating smoothed optical flow [41] between I'_{t-1} and I_t, independently within each region. Note that we do not assume or require accurate motion estimation at this stage. Fig. 6 (bottom-left) provides an example region map from the BOY sequence $t-1$ warped according to motion field $I'_{t-1} \rightarrow I_t$.

We select a subset of the motion propagated pixels (written $\mathcal{O}$), and their corresponding region assignments, as prior labels to influence the segmentation of I_t. To mitigate the impact of imprecise motion estimation, we form $\mathcal{O}$ by sampling from a morphologically thinned skeleton of the motion propagated regions (Fig. 6, bottom-right). This approach is inspired by the "scribbles" used in the interactive GrabCut system [36], but note that we perform an automatic and multi-region (as opposed to binary) labeling. The skeleton emphasizes geometrical and topological properties of the motion propagated region map, such as its connectivity, topology, length, direction, and width. To further deal with the uncertainties in positions which are closer to the estimated region boundary, we use only the skeletons whose distance to the boundary exceeds a pre-set confidence. Fig. 6 illustrates the complete process, which we find to be tolerant to moderate misalignments caused by inaccurate motion estimation.

We build a GMM color model for each label l_i, sampling the historical colors of labeled pixels over recent frames. To cope with luminance variations in the sequence, the proportion of samples $S_{l_i,t-d} \in [0,1]$ ($d > 0$) drawn from all l_i-labeled pixels from historical frame I_{t-d} decreases exponentially as the temporal distance d from the current frame I_t increases (Fig. 7):

$$S_{l_i,t-d} \propto e^{-d^2/\sigma_d^2}. \tag{9}$$

Our system selects a smaller σ_d when luminance variance is large, contributing more recent data to the GMM, otherwise the historical data contributes more to increase robustness.

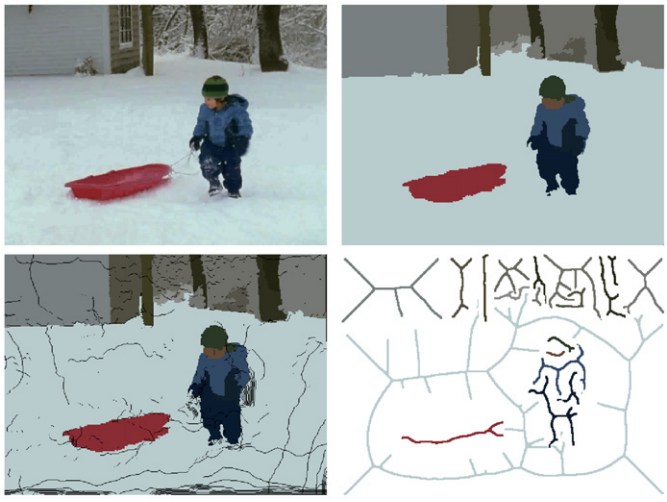

Fig. 6. Prior propagation: (Top-left) video frame I_{t-1}; (Top-right) region labeling of I_{t-1} following multi-label graph cut; (Bottom-left) region labels warped according to per-pixel motion flow field $I'_{t-1} \rightarrow I_t$—for example, note the shift of the boy's left glove. (Bottom-right) Thinning yields prior labels for the segmentation of I_t.

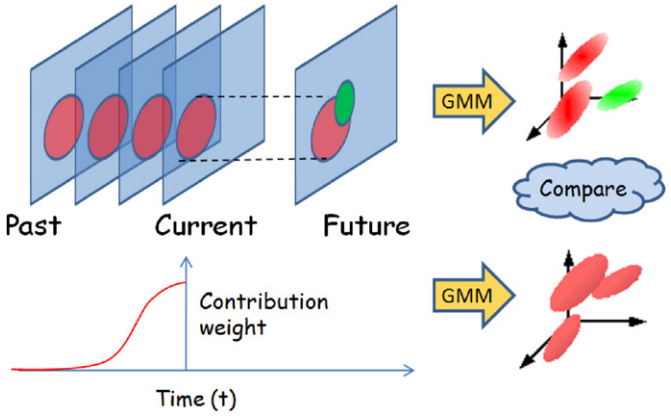

Fig. 7. A GMM color model of each region is built incrementally over time, with contributions biased toward more recent observations. If the GMM of a region abruptly changes color distribution (χ^2 metric) then the region is re-segmented (Section 4.3).

4.3. Refining region labels

The method of Section 4.1 labels I_t with some or all of the region labels in use in the region map at $t-1$. However, new objects may appear in the sequence over time I_t due to occlusion effects of objects moving into shot. This is most apparent in clips such as DRAMA (Fig. 14). These objects may warrant introduction of a new region label, should they differ in color from existing regions. In such a situation, pixels comprising the object are erroneously labeled from the existing label set by the graph-cut optimization, which in turn perturbs the color distribution of the region. We can detect this by measuring the χ^2 distance (as defined by Hall and Hicks [42]) between the GMM of a region at time t and the historical GMM built over time (Fig. 7).

For successive frames, we keep two sets of color models for each label l in frame I_t being processed: (1) Historical color models associated with each label $M^h_{l:\{l \in \mathcal{L}\}} \coloneqq G_l(I_{t-4}, I_{t-3})$ and (2) an updated color model $M^u_{l:\{l \in \mathcal{L}\}} \coloneqq G_l(I_t)$. We set a guard interval of two frames (I_{t-1} and I_{t-2}) between those two models to detect a significant change. If the χ^2 distance between these two models exceeds a threshold, new objects are deemed present.

To build color models for the new objects we extract the dominant modes of colors within the region. We apply mean-shift to perform unsupervised clustering on the spatial-color modes (XY+RGB) of pixels in the region. This yields a localized segmentation of pixels in the region. We extend our label set to accommodate each new region arising from the mean-shift segmentation, and for each new region also compute GMM color models and region skeletons as in Section 4.2. Re-applying the graph-cut optimization locally within the region, using these new labels and constraints, yields an improved segmentation for I_t that is carried to successive frames.

4.4. Smoothing and filtering

Our segmentation algorithm produces stable region maps, but due to visual ambiguities in poor contrast areas, the location of region boundaries tend to oscillate in position by a few pixels. We can attenuate this effect by performing spatio-temporal smoothing. Specifically, by coherently labeling regions in adjacent frames, we have formed a set of space-time volumes. Applying a fine scale ($3 \times 3 \times 3$) Gaussian filter removes boundary noise. We avoid removing detail by only filtering volumes above a certain size.

We inspect the duration $d_{l,k}$ of the disconnected space-time volume k ($k = 1, \ldots, K_{obj_l}$) with the same label l, in a time window of

24 frames (1 s). If the duration of any of these disconnected video object within this time window is shorter than a length

$$D_{l:\{1,\ldots,L\}} = \min\left\{\max_{k\in\{1,\ldots,K_{obj_l}\}} d_{l,k}, \tau_r\right\}, \quad (10)$$

this space-time volume is removed. τ_r is set to be six frames (about $\frac{1}{4}$ s). The effect of this process is that the spurious volumes due to false segmentation and short-lived objects are removed, as shown in Fig. 8. The "holes" left by filtering and smoothing are filled by extrapolating region labels from immediate space-time neighbors on a nearest-neighbor basis.

4.5. Stroke placement and shading

Our video segmentation algorithm ensures regions not only deform in a coherent manner, but are also labeled consistently between frames. This space-time description of scene structure may be rendered in a variety of artistic styles; here we give an example of one shading and one stroke based style.

4.5.1. Cartooning

Superimposing black edges over regions shaded with their mean pixel color can produce coherent cartoon effects (Fig. 11). In our cartoon examples, a mask of inter-region boundaries is produced for each frame. We identify "junction" points on region boundaries by identifying 3×3 pixel windows containing >2 region labels—and remove the corresponding boundary fragments from the mask. This results in a series of connected pixel chains that we transform into $\beta-$spline strokes by sampling knots at equi-distant intervals. The strokes are rendered as dark brush strokes, with thickness proportional to stroke length (after Wang et al. [21]) tapering toward the stroke ends. We render frames independently without further post-processing; this is both for simplicity and to demonstrate the temporal coherence of our segmentation output.

We can also exploit the temporally corresponded region labeling to differentially render regions of interest. For instance, users are particularly sensitive to over-abstraction of detail in faces; commonly present in home video footage. We run human face detection [43] over frames to identify labeled regions likely to contain faces. Internal detail in these regions may be restored by blending in a posterized image of underlying video footage, and detail further enhanced by subtracting a Laplacian of Gaussian (LoG) filtered image from the result.

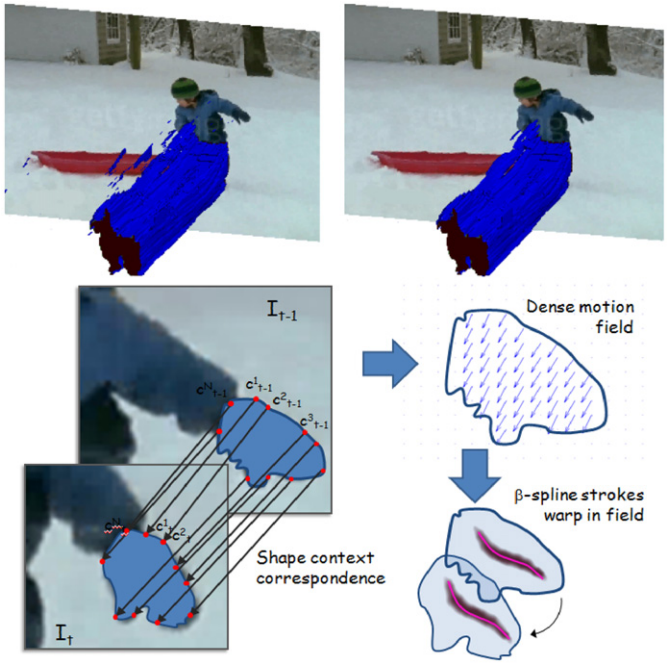

Fig. 8. Above: co-labeled regions are smoothed in space-time to remove any spurious regions. Below: Brush strokes are painted on a stable reference frame, created by corresponding co-labeled regions in adjacent frames and interpolating a dense motion field.

4.5.2. Painterly rendering

Alternatively we can paint $\beta-$spline brush strokes inside regions, coherently deforming those splines by warping their control points to match the motion of the region boundary (similar to the manually bootstraped rotoscoping system presented by Agarwala et al. [44]). Boundary correspondences are computed between temporally adjacent, co-labeled regions using Shape Contexts [45]. The set of N corresponded boundary locations $\{\langle c^1_{t-1},c^1_t\rangle,\langle c^2_{t-1},c^2_t\rangle,\ldots,\langle c^N_{t-1},c^N_t\rangle\}$ is used to derive the motion vector for a control point p at time t as

$$p = \frac{1}{N}\sum_{i=1}^{N}\omega(p,c^i_t)|c^i_t - c^i_{t-1}|, \quad (11)$$

where $\omega(\cdot)$ is a Gaussian weighted function of the shortest distance between two points within the region (see Fig. 8, below). Our coherent segmentation promotes smooth deformation of region shape, and so flicker-free motion of brush strokes.

We paint the $\beta-$spline strokes within a region using Hertzmann's bi-directional stroke growth algorithm [46]. In the original algorithm, strokes are grown from random seed points using the orientation of an intensity gradient field computed from the underlying image. However, computing such orientation directly from video footage typically promotes incoherence. Instead, we interpolate an orientation field from the shape of the region. Orientations are locally obtained at points of correspondence on the boundary $\theta[x,y]\mapsto \mathrm{atan}(c^{i-1}_t - c^i_t)$. We define a dense orientation field Θ_Ω over all coordinates within the region $\Omega\in\mathbb{R}^2$, minimizing:

$$\underset{\Theta}{\mathrm{argmin}}\iint_\Omega(\nabla\Theta-\mathbf{v})^2 \quad \text{s.t. } \Theta|_{\delta\Omega}=\theta|_{\delta\Omega}. \quad (12)$$

i.e. $\Delta\Theta=0$ over Ω s.t. $\Theta|_{\delta\Omega}=\theta|_{\delta\Omega}$ for which a discrete solution was presented by Perez et al. [47] solving Poisson's equation with Dirichlet boundary conditions. $\mathbf{v}$ represents the first order derivative of θ. Examples of painterly output are given in Fig. 15.

5. Content sequencing

Finally, we explain the algorithms for sequencing stylized content to create the temporal composition of media items from the user's collection, and for creating the animated transitions between displayed items.

5.1. Temporal composition

We desire the DAD to autonomously transition between a sparse yet diverse sample of user content to present a summary of the collection. "Sparse" describes a small number of representative images from each semantic category. This is achieved using the *hierarchical representation* of the media collection constructed during pre-processing (Section 3).

Recall each node in our representation encodes a *cluster* of similar media; defined using either a semantic similarity measure (toward the root) or a visual similarity measure (toward the leaves). For each node in the hierarchy, similarities between all media items within the cluster are computed (Section 3.1), to form a new graph of media items—with edge weights indicating the (dis-)similarity of a pair of items. Computing the shortest Hamilton

cycle within this graph creates a non-repetitive set of transitions maximizing the similarity between successive media items and thus the coherence of the sequence. Although a precise solution maps to the classical NP-hard "traveling salesman problem" (TSP), an approximate solution can be found quickly using heuristic search methods. We adopt a Genetic Algorithm (GA) based solution [48]. In practice TSP paths for each cluster are pre-processed.

The DAD is equipped with a camera and a face detection technology [43] to detect user gaze (attention) directed towards the display (Fig. 12). Sequencing proceeds in one of the two modes, depending on whether user attention is present or not.

When attention is not present, the intention is to create a succinct and diverse summary of the collection and speculatively display content that might catch a passing user's interest. Clusters in the first level of the tree represent coarse semantic categories across the whole collection. Displaying a sample of media from each cluster in turn yields such high-level summary of the collection. However, rather than present a random succession of media, we desire a degree of temporal coherence in our choice of media to create compelling paths through the collection—to tell a "story" through visual media, with the aim of prompting user reminiscence. Although TSP paths within a cluster offer coherence for intra-cluster transitions, they do not offer inter-cluster coherence. We address the latter by identifying a "semantic route" between media items (A,B) in different clusters. This is achieved by taking the union of media clusters containing A and B, and computing a (dis-)similarity graph as before. Dijsktra's algorithm yields a shortest path between A and B, encoding the "semantic route"; the most coherent sequence of media items to transition between A and B (Fig. 9).

When attention is present, we do not permit jumps between siblings, but instead permit transitions to the parent or children of the current cluster. These transitions represent "generalization" or "drilling down" into a media topic, respectively. Suppose we have detected user interest in a media item A. We stochastically choose to either remain on the TSP path containing A in the current cluster, or to transition to a child cluster also containing A (i.e. begin transitioning along the TSP path in that cluster). For our experiments the probability of continuing on the current path, or "drilling down" is even. When interest in the display abates, we transition back "up" the tree in a similar manner; with even choice between continuing on the TSP path or jumping to the parent TSP path. In both cases A exists within the destination cluster, and so sequencing continues simply by switching the display process onto the TSP path for the new cluster to ensure smooth transitions when transitioning up and down the tree.

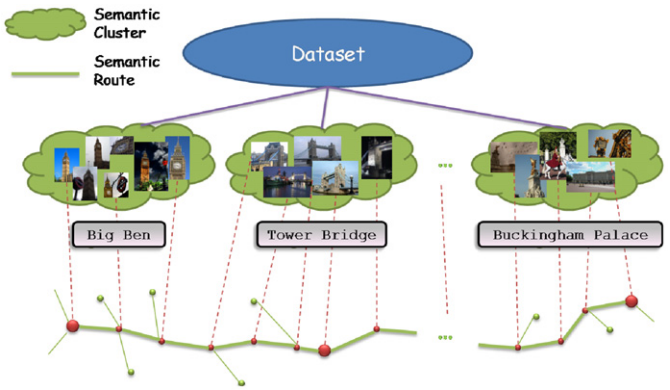

Fig. 9. Transitions are made across top-level clusters by sequencing display of content along "semantic route" between a source and destination media item (depicted as large red nodes). The semantic route is the shortest path computed across the graph; here nodes part of that route are shaded red, otherwise green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Rendering transitions

Having established a sequencing mechanism during visualization, we animate the transition between stylized media items according to the scene structure (region map).

We first establish a mapping between each region R^j_{t-1} and R^i_t corresponding, respectively, to the final and initial frames of the two clips (recall that images are accommodated in our framework as single frame videos). The region mapping is created in a greedy manner, iteratively pairing off regions that minimize:

$$\underset{\{i,j\}}{\operatorname{argmin}}[w_1 \;\; w_2 \;\; w_3]\begin{bmatrix} C(R^j_{t-1},R^i_t) \\ A(R^j_{t-1},R^i_t) \\ S(R^j_{t-1},R^i_t) \end{bmatrix}, \tag{13}$$

where the normalized functions: $C(\cdot)$ indicates mean color similarity; $A(\cdot)$ indicates relative area; $S(\cdot)$ indicates shape similarity in terms of region compactness. We bias weights ω_{1-3} empirically to 0.5, 0.4, 0.1. The greedy assignment continues until (13) falls below a threshold. Unassigned regions in the mapping are animated to "disappear" (shrink to a point at the centroid) or "appear" (grow from the centroid); whereas regions mapped between frames are animated to morph into one another.

Regions are morphed using simple linear blending. Each region is vectorized into a polygon and a series of regularly spaced control vertices established on the boundary. A correspondence is established between vertices of R^j_{t-1} and R^i_t to minimize distance between corresponded vertices. The position of control vertices are linearly blended over time (typically $\frac{1}{4}$ s) to animate the region from one shape to another. Region color is similarly blended. Although more complex vertex correspondence approaches were investigated [45], these lacked stability when presented with moderate changes in region shape. The resulting transitions are shown in Fig. 10.

6. Results and user study

We present a qualitative comparison of the proposed video segmentation algorithm with two existing techniques [33,49] and present a gallery of stills from videos stylized into cartoons and paintings. We also present a small-scale study exploring user engagement with the DAD.

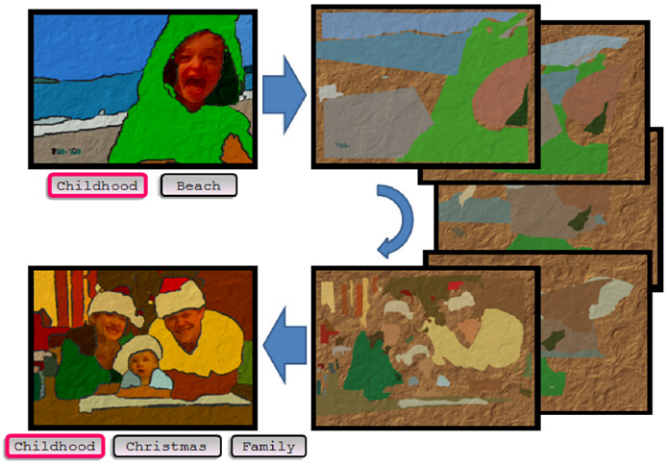

Fig. 10. Frames from the transition animation between two clips.

6.1. Video segmentation and stylization

To demonstrate the advantages of the proposed multi-label video segmentation algorithm, we compare the approach proposed in Section 4 to two leading segmentation methods for per-frame [33] and spatio-temporal [49] segmentation (Fig. 13). We observe the region boundaries in the proposed method to exhibit improved stability over time. Fig. 14 indicates the region maps produced by the segmentation algorithm over four video sequences. We test our algorithm on fast moving footage containing small objects ("BEAR" from Collomosse et al. [12]). Unlike previous work, fine scale features (e.g. the bear's eyes and nose) are retained. Similarly, "DANCE" demonstrates the ability to cope with fast motion and partial occlusions. "DRAMA" shows correct handling of regions that disappear and appear within sequences, the latter detected by changes in the region color distribution and addressed as out-lined in Section 4.3. The "KITE" sequences show the aesthetic ability to selectively abstract detail (trees) from the stylized video, when interactively removed by the user in the initial frame. In all cases our segmentations appear flicker-free; some flicker is occasionally present at the bottom-left of clips due to the frame identifier which could be manually abstracted away by similarly modifying the initial frame.

We demonstrate the video stylization and transition animations using a collection of 23 videos. Fig. 11 shows representative frames of the stylized footage in both cartoon and painterly styles; 6 min of the perpetual animated display is also included in the supplementary material. An example transition animation is given in Fig. 10. The resulting clip transition animations match large, similarly colored regions between frames producing a pleasing smooth transition effect evident throughout the DAD sequence.

Following coherent segmentation, Fig. 15 shows frames of painterly renderings over four video sequences in natural scenes. The smooth deformation of regions enables stable and flicker-free motion of brush strokes, which produces an aesthetically pleasing painterly effect over the input video sequences.

6.2. Study of user engagement

We evaluated the efficacy of our content sequencing algorithm (Section 5.1) in a small-scale user study. Our hypothesis was that the DAD media sequencing algorithm would prove more engaging for users than simple random slideshows, as typified by existing commercial digital photo frames. We tested the algorithm with and without adaptation to user interest. We measured user engagement using the DAD's gaze detection; counting the proportion of displayed media items that attracted user attention. In this evaluation we used photographic media only, rendered in a painterly style (Section 4.5.2). We eschewed video content to potential bias introduced by movement which can act as a strong attractor of attention. Transition animations were also disabled.

6.2.1. Experimental setup

The media collection comprised 600 user-tagged Flickr images of 18 landmarks in London, licensed under the Creative Commons. The dataset is adopted to emulate a typical home media collection consisting of photos taken from a tour in London, which also facilitates the user study. We de-noised the associated tags by stripping numeric tags, any punctuation and commonly used Flickr tags that do not relate to content e.g. camera model. The semantic relevance between the tags was pre-computed, and images analyzed to form a BoW code-book with 4000 visual words. The study comprised 10 participants between the ages of 20–40, of mixed gender, with varying levels of technical expertise. The DAD device (1.33 GHz CPU, 512 MB RAM, 800 × 480 resolution) was positioned within proximity to the participant in their everyday working environment. The media collection is stylized beforehand and stored in the DAD. The DAD performs face detection, content sequencing and visualization. The camera and face detector were calibrated to record an attention event when the user's head is oriented towards the DAD.

Fig. 12. One of the three identical DAD devices used in the experiment. The DAD adapts in real time to display interesting contents at various levels of detail in response to user attention level.

Fig. 11. A collage of stylized frames sampled from the user video collection studied in this paper.

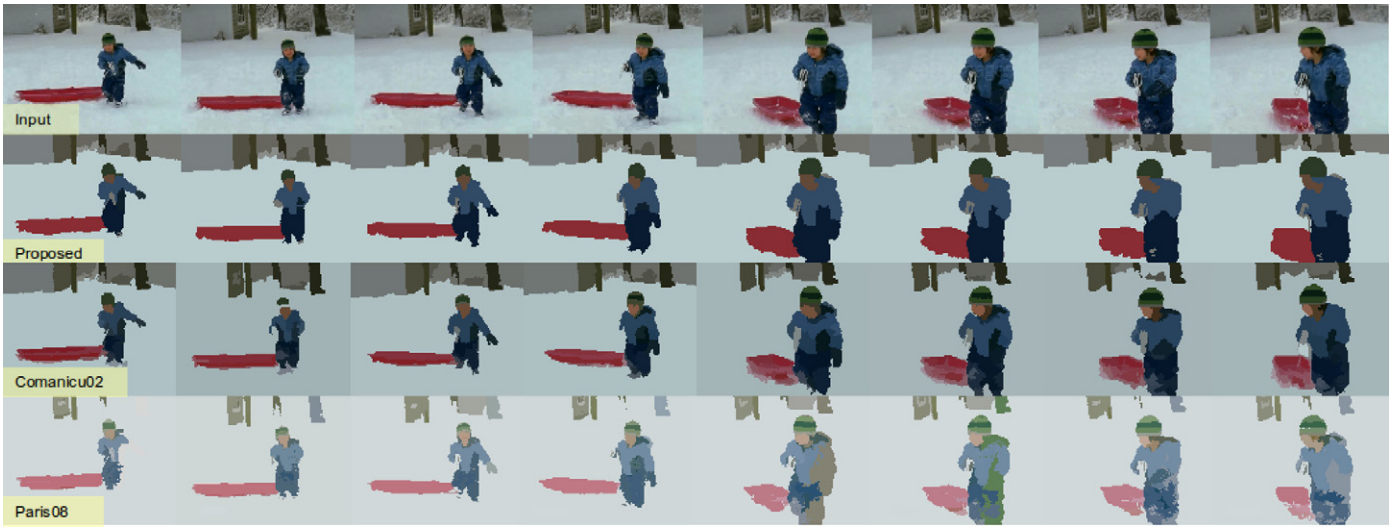

Fig. 13. Comparing the accuracy and coherence of our segmentation algorithm on the BOY sequence, to "synergistic" mean-shift + edge [33] and a state of the art spatio-temporal method [49]. Boundaries are less prone to variation in shape and topology.

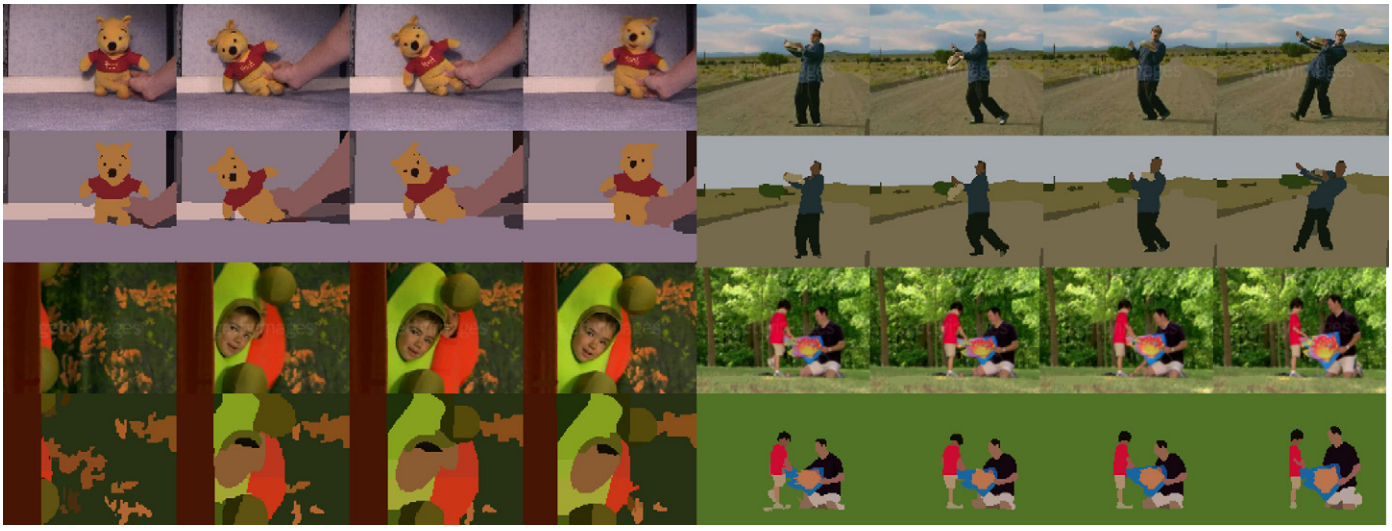

Fig. 14. Illustrating the coherent region maps produced by our segmentation method. Top: BEAR and DANCE contain small regions moving quickly over time. Bottom: The DRAMA sequence shows correct handling of regions appearance. The KITE sequence indicates how background detail may (optionally) be abstracted by modifying the initial frame segmentation to merge unwanted detailed regions.

6.2.2. Protocol

The experiment comparatively evaluates three operational modes of the DAD device:

- Random and passive display (RP): the DAD randomly selects non-repetitive images from the dataset to display.
- Structured and passive display (SP): the DAD displays images in the proposed approach without responding to user attention.
- Structured and adaptive display (SA): the DAD displays images in the proposed approach.

Participants might be involved in parallel activities while seated during the experiments and only pay extended attention to the screen when attracted by the content. In order to mitigate the effect of short term memory of the image collection we shuffle the order in which the three modes are evaluated by each user. Furthermore, during the evaluation, the participants were unaware of the nature of the three modes or the interactive nature of mode SA. All user attention events during the three presentations for each user are recorded in the background for analysis. Experiment for each mode takes 20 min and each image is shown for 3.5 s.

6.2.3. Experimental results and feedback

Attention events recorded by the DAD for the three operational modes are given in Table 2. Table 1 records the proportion of images that attracted user attention out of the total images displayed for each mode. A paired t-test between pairings of modes indicate strong statistical significance between each mode of operation (Table 2). This suggests a qualitative improvement in level of user engagement using our structured sequencing approach, versus a random slideshow. Improvement is also observed with user adaptive sequencing (SA) over non-adaptive (SP). Quantifying this improvement by averaging across the users the SA mode we record $\sim 17\%$ more attended images than the SP scheme, which in turn recorded $\sim 55\%$ more attended images than the RP scheme. Data on how user attention was distributed across the semantic clusters in the SA case shows large differences between the clusters, consistent with the adaption strategy adopted i.e. the participant's

Fig. 15. Examples of coherent painterly renderings produced from the BOY, KITE, PICNIC and DANCE videos (top to bottom).

Table 1
Proportion of images attracting user attention out of total displayed images in each scheme.

	User #1	User #2	User #3	User #4	User #5	User #6	User #7	User #8	User #9	User #10
RP	0.0467	0.1133	0.1400	0.1050	0.0750	0.1350	0.0850	0.0367	0.0300	0.1433
SP	0.0850	0.1783	0.2033	0.0783	0.1300	0.1767	0.1483	0.0833	0.0950	0.2383
SA	0.1083	0.2333	0.1750	0.1017	0.1783	0.2000	0.1800	0.1250	0.1267	0.2250

initial interest in a specific category is detected and draws more images from that category that serve to maintain that interest. We thus conclude that our proposed approach offers an engaging means to display the contents of a large media collection with minimal user interaction.

During a questionnaire based de-brief, user feedback on the DAD was broadly positive, both in terms of the aesthetics of the painterly rendering and the DAD system. Participants feel engaged with the system, remarking on the structured and adaptive manner of presentation. Eighty percent of the participants find that the DAD

Table 2
Significance of results (paired *t*-test).

t-Test	Means	Std. dev.	*p*-Value
RP/SP	0.0910/0.1417	0.0431/0.0565	0.0007
RP/SA	0.0910/0.1653	0.0431/0.0474	0.0001
SA/SP	0.1653/0.1417	0.0474/ 0.0565	0.0184

displays more images of interest in our proposed approach (SA) than in the other two approaches and regard it as a useful means to display their own digital media collections. Seventy percent of the participants deem that our proposed approach (SA) presents an effective global summary of the structure of the collection. Sixty percent of the participants would consider presenting their own digital media collections in a similar painting style, with the remainder concerned about the recognition of faces in the stylized content. Sixty percent of participants would like to be able to take over control of the presentation. Suggested controls are: (1) Ban or skip a specific category. (2) Hold on the content being displayed. (3) Alternative artistic rendering styles. (4) Indication of user attention being detected. All participants are satisfied with the hardware specifications of the DAD device, such as the appearance, screen size, screen brightness, and speed. Participants filled in the subjective questionnaire without knowing the DAD mode they were commenting on.

7. Conclusion

We have presented a *digital ambient display* (DAD) that harnesses artistic stylization to create an abstraction of user's experiences through their home digital media collections. The DAD automatically selects, stylizes and transitions between media contents enabling users to passively or actively consume their digital media collections and rediscover past memories.

We contributed a novel algorithm for coherent video segmentation based on multi-label graph cut, and applied this algorithm to stylized animation in the DAD. By parsing the video into coherent spatial segments, we are able to represent scene structure. This representation allows us to establish correspondence between frames, enabling the coherent stylization of space-time volumes with both shading and painterly effects. The latter was possible by painting brush strokes on a smoothly deforming reference frame defined by the regions. We are also able to create aesthetically pleasing transition effects between different video clips using region correspondence. Video segmentation could be further enhanced by exploring the backward propagation of region labels to further improve coherence of segmentation. We would also like to improve the painterly rendering by differentiating between region motion caused by occlusion vs. object deformation, to more closely align the movement of painted strokes to the perceived structure in the scene.

A further contribution of the paper is a novel approach to structuring and navigating visual media collections. We described an algorithm for adaptively sequencing media items using graph optimization in a coarse-to-fine manner driven by user attention. By recursively clustering media items into a hierarchy, we were able to plan routes within clusters to display content of a common theme. We were also able to plan routes between clusters to summarize media within the collection. We deployed our system on dedicated hardware and undertook a small-scale user trial to validate our content sequencing algorithm, which was shown to be more engaging than random photo slideshows.

In future work we would like to offer more control to the user over presentation. An improved interface might enable users to ban or skip specific categories they are less interested in, and hold on interesting contents for closer inspection. In addition to global visual similarity of media items (addressed in this paper) it might be interesting to harness recent developments in image cosegmentation [2,50] to enable users to explore "similar content" within a region of interest indicated by touching a particular area on the display.

Acknowledgement

This work was funded by Hewlett Packard under the IRP studentship programme (Grant #477).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version of 10.1016/j.cag.2010.11.004.

References

[1] Combs TTA, Bederson BB. Does zooming improve image browsing? In: DL '99: Proceedings of the fourth ACM conference on digital libraries. ACM; 1999. p. 130–7.
[2] Heath K, Gelfand N, Ovsjanikov M, Aanjaneya M, Guibas LJ. Image webs: computing and exploiting connectivity in image collections. In: CVPR, 2010.
[3] Rodden K, Basalaj W, Sinclair D, Wood K. Does organisation by similarity assist image browsing? In: CHI '01: Proceedings of the SIGCHI conference on human factors in computing systems. New York, NY, USA: ACM; 2001. p. 190–7.
[4] Schaefer G. A next generation browsing environment for large image repositories. Multimedia Tools and Applications 2010;47(1):105–20.
[5] Chen J, Bouman C, Dalton JC. Hierarchical browsing and search of large image databases. IEEE Transactions on Image Processing 2000;9:442–55.
[6] Goldberger J, Gordon S, Greenspan H. Unsupervised image-set clustering using an information theoretic framework. IEEE Transactions on Image Processing 2006;15:449–58.
[7] Krishnamachari S, Abdel-Mottaleb M. Image browsing using hierarchical clustering. In: IEEE symposium on computers and communications, 1999. p. 301–7.
[8] Arksey N. Exploring the design space for concurrent use of personal and large displays for in-home collaboration. Master's thesis, University of British Columbia; August 2007.
[9] You W, Feis S, Lea R. Studying vision-based multiple-user interaction with in-home large displays. In: Proceedings of the 3rd ACM workshop on human-centred computing (HCC), 2008. p. 19–26.
[10] Wang T, Collomosse J, Slatter D, Cheatle P, Greig D. Video stylization for digital ambient displays of home movies. In: Proceedings of the ACM NPAR, 2010. p. 137–46.
[11] DeCarlo D, Santella A. Stylization and abstraction of photographs. In: SIGGRAPH '02: Proceedings of the 29th annual conference on computer graphics and interactive techniques. ACM; 2002. p. 769–76.
[12] Collomosse J, Rowntree D, Hall P. Stroke surfaces: temporally coherent artistic animations from video. Transactions on Visualization and Computer Graphics 2005;11:540–9.
[13] Schodl A, Skeliski R, Salesin D, Essa H. Video textures. In: Proceedings of the ACM SIGGRAPH, 2000. p. 489–98.
[14] Kovar L, Gleicher M, Pighin F. Motion graphs. In: Proceedings of the ACM SIGGRAPH, 2002. p. 473–82.
[15] Sivic J, Zisserman A. Video google: a text retrieval approach to object matching in videos. In: Proceedings of the international conference on computer vision, vol. 2, 2003. p. 1470–7.
[16] Litwinowicz P. Processing images and video for an impressionist effect. In: SIGGRAPH, 1997. p. 407–14.
[17] Hays J, Essa IA. Image and video based painterly animation. In: NPAR, 2004. p. 113–20.
[18] Hertzmann A, Perlin K. Painterly rendering for video and interaction. In: NPAR, 2000. p. 7–12.
[19] Meier BJ. Painterly rendering for animation. In: Proceedings of the ACM SIGRGAPH, 1996. p. 477–84.
[20] Collomosse J. Higher level techniques for the artistic rendering of images and video. PhD thesis, University of Bath; May 2004.
[21] Wang J, Xu Y, Shum H, Cohen M. Video tooning. In: SIGGRAPH, vol. 23, 2004. p. 574–83.
[22] Winnemoller H, Olsen S, Gooch B. Real-time video abstraction. In: ACM SIGGRAPH, 2006. p. 1221–6.
[23] Kyprianidis JE, Döllner J. Image abstraction by structure adaptive filtering. In: Proceedings of the EG UK theory and practice of computer graphics, 2008. p. 51–8.
[24] Kang H, Lee S, Chui CK. Flow-based image abstraction. IEEE Transactions on Visualization and Computer Graphics 2009;15(1):62–76.

[25] Kyprianidis J-E, Kang H, Doellner J. Image and video abstraction by anisotropic Kuwahara filtering. In: Proceedings of the Pacific graphics, vol. 28, 2009.
[26] Bhat P, Zitnick CL, Cohen M, Curless B. Gradientshop: a gradient-domain optimization framework for image and video filtering. ACM Transactions on Graphics 2010;29(2):1–14. doi:http://doi.acm.org/10.1145/1731047.1731048.
[27] Bai X, Wang J, Simons D, Saprio G. Video snapcut: robust video object cutout using localized classifiers. In: ACM SIGGRAPH, 2009.
[28] Price B, Morse B, Cohen S. Livecut: learning-based interactive video segmentation by evaluation of multiple propagated cues. In: ICCV, 2009.
[29] Wang T, Mansfield A, Hu R, Collomosse J. An evolutionary approach to automatic video editing. In: Proceedings of the 6th European conference on visual media production (CVMP), 2009.
[30] Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;315:972–6.
[31] Sigurbjörnsson B, van Zwol R. Flickr tag recommendation based on collective knowledge. In: WWW 2008, 2008. p. 327–36.
[32] van de Sande KEA, Gevers T, Snoek CGM. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2010;32(9):1582–96.
[33] Comaniciu D, Meer P. Mean shift: a robust approach toward feature analysis. Transactions on Pattern Analysis and Machine Intelligence 2002;24:603–19.
[34] Rother C, Kolmogorov V, Blake A. Grabcut: interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics 2004;23:309–14.
[35] Boykov Y, Funka-Lea G. Graph cuts and efficient n-d image segmentation. International Journal of Computer Vision 2006;2(70):109–31.
[36] Blake A, Rother C, Brown M, Prez P, Torr P. Interactive image segmentation using an adaptive GMMRF model. In: ECCV, 2004. p. 428–41.
[37] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. Transactions on Pattern Analysis and Machine Intelligence 2001;23:1222–39.
[38] Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. Transactions on Pattern Analysis and Machine Intelligence 2004;26:1124–37.
[39] Alahari K, Kohli P, Torr, Reduce, reuse & recycle: efficiently solving multi-label MRFs. In: CVPR, 2008. p. 1–8.
[40] Lowe D. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2004;60:91–110.
[41] Black M, Anandan P. A framework for the robust estimation of optical flow. In: ICCV, 1993. p. 231–36.
[42] Hall PM, Hicks Y. CSBU-2004-03: a method to add gaussian mixture models. Technical Report, University of Bath, 2004.
[43] Viola P, Jones M. Robust real-time object detection. International Journal of Computer Vision 2004;57(2):137–54.
[44] Agarwala A, Hertzmann A, Salesin D, Seitz S. Keyframe-based tracking for rotoscoping and animation. In: Proceedings of the ACM SIGGRAPH, 2004. p. 294–302.
[45] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. Transactions on Pattern Analysis and Machine Intelligence 2002;24:509–21.
[46] Hertzmann A. Painterly rendering with curved brush strokes of multiple sizes. In: Proceedings of the ACM SIGGRAPH, 1998. p. 453–60.
[47] Perez P, Gangnet M, Blake A. Poisson image editing. In: ACM SIGGRAPH, 2003. p. 313–8.
[48] Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley; 1989.
[49] Paris S. Edge-preserving smoothing and mean-shift segmentation of video streams. In: ECCV, 2008. p. 460–73.
[50] Jan Čech MP, Matas J. Efficient sequential correspondence selection by cosegmentation. In: CVPR, 2008.